**Java课程项目设计文档**

1、程序名：Github托管项目爬虫

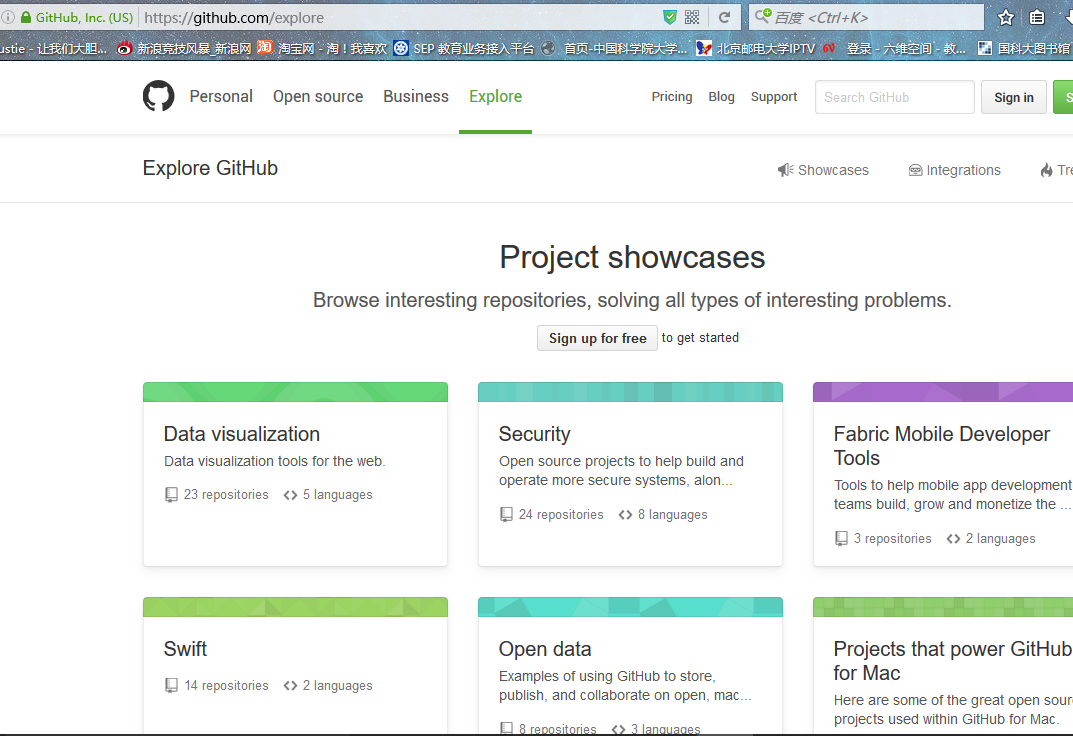
2、开发人员：

2015E8009361055 陈思远

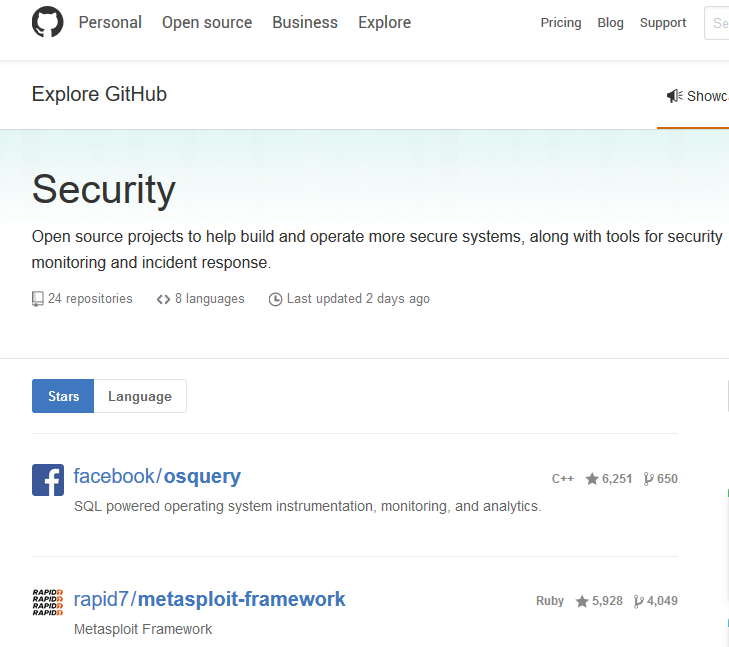
2015E8009361053 张涵笑

3、程序功能

本程序在webmagic开源爬虫框架的基础上，自定义实现了抓取网址<https://github.com/explore>（使用程序前请确保主机能连接到该网址）上的Github托管开源项目的大致信息。网址截图如下图所示：



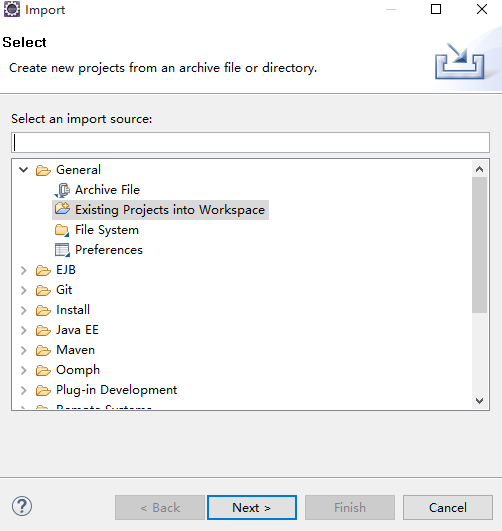
打开其中任何一个project case，例如Security，截图如下图所示：



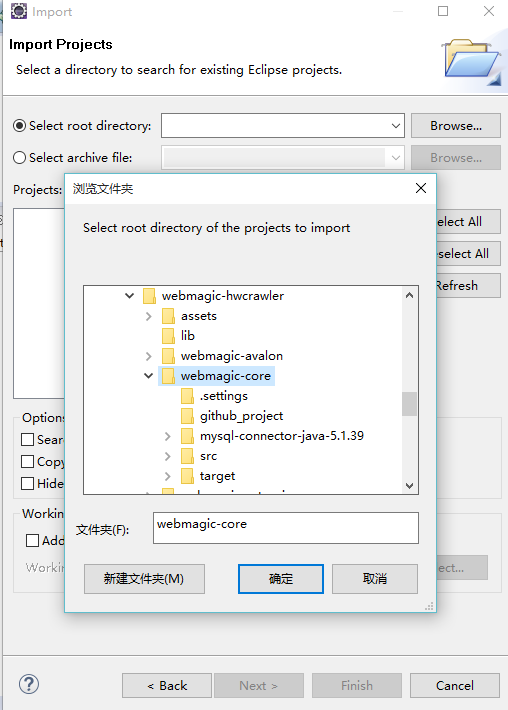
程序会抓取该页面上所有的项目的信息，包括项目名，开发语言，标星次数，fork次数以及项目简要描述。例如facebook / osquery，C++ ，6,251， 650，SQL powered operating system instrumentation, monitoring, and analytics.便是对项目facebook / osquery所抓取的数据。

4、项目安装及运行说明：

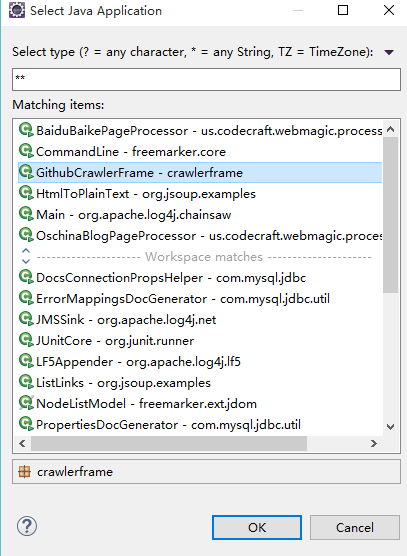
在eclipse上导入工程，点击File->import,选择General->Existing projects into workspace,如下图所示：



接着点击Next，在select root directory选中webmagichwcrawler下的webmagic-core目录，点击确定，Finish，即导入工程成功。如下图所示：

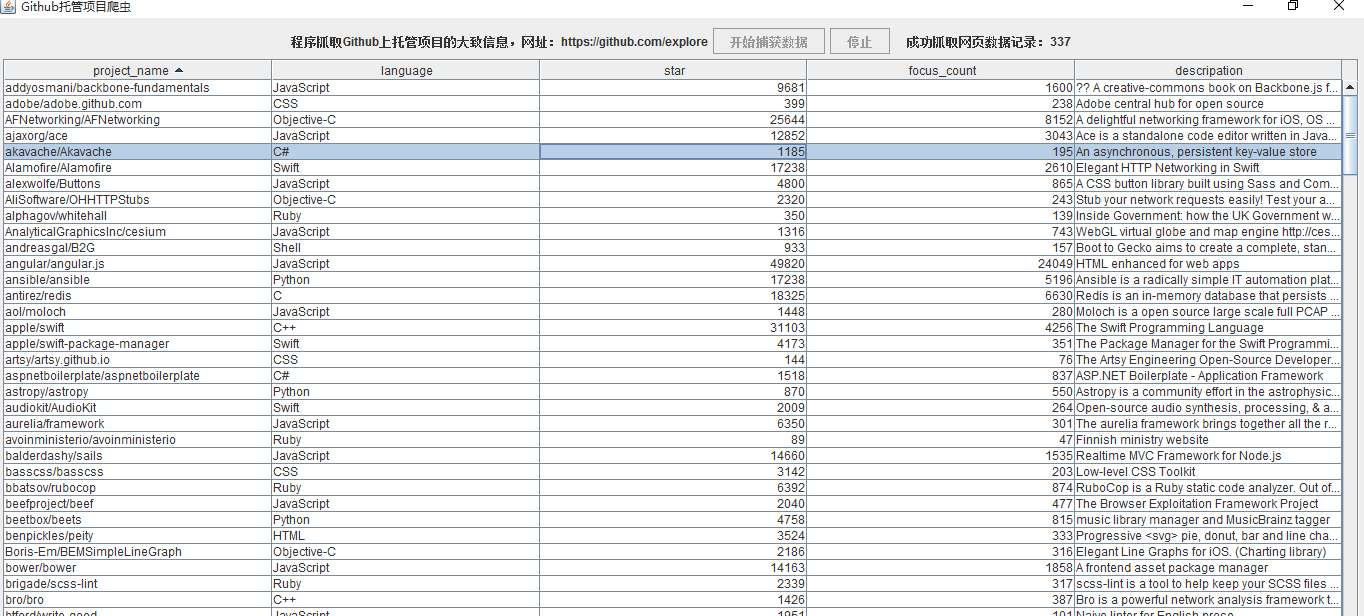


右键单机项目名webmagic-core，点击run as，选择java application，选择GithubCrawlerFrame，如下图所示：

程序启动GUI界面如下图所示：



点击开始捕获数据按钮后，等待大约30s时间，界面停止按钮可见，点击后可得结果界面，如下图所示：



由图可知Github网站上托管了337个开源项目，点击表格的列名，可以按所选列升序或降序排列。

5、开发人员分工：

陈思远的工作是负责抓取网页数据，设计数据库结构及操纵函数。抓取网页数据主要包括建立与目标网页的Socket连接，抓取符合要求的网页链接，从网页所返回的HTML报文中解析出需要的字段。数据库设计主要包括数据库表的设计及实现，同时也要给出数据库操纵的函数定义。

张涵笑的工作是负责GUI界面的设计。主要包括表格的设计，对显示数据的逻辑控制（读取数据库数据并显示到界面表格中），数据的统计和列排序功能的实现。